

AD-A128 645

EFFECTS OF STANDARD EXTREMITY ON MIXED STANDARD SCALE  
PERFORMANCE RATINGS. (U) PURDUE UNIV LAFAYETTE IN DEPT  
OF PSYCHOLOGICAL SCIENCES J L BARNES-FARRELL ET AL.

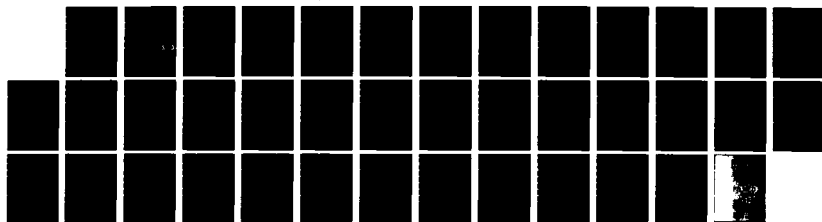
1/1

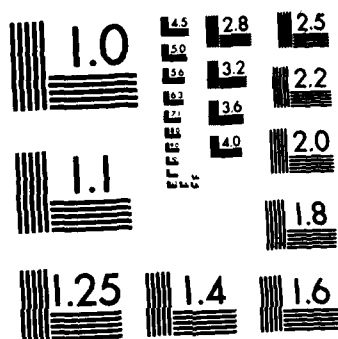
UNCLASSIFIED

MAR 83 83-2 N00014-82-K-0449

F/G 5/10

NL





MICROCOPY RESOLUTION TEST CHART  
NATIONAL BUREAU OF STANDARDS-1963-A

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

12

## REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS  
BEFORE COMPLETING FORM

1. REPORT NUMBER 83-2	2. GOVT ACCESSION NO. AD-A128645	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Effects of Standard Extremity on Mixed Standard Scale Performance Ratings		5. TYPE OF REPORT & PERIOD COVERED Interim
7. AUTHOR(s) Janet L. Barnes-Farrell and Howard M. Weiss		6. PERFORMING ORG. REPORT NUMBER
9. PERFORMING ORGANIZATION NAME AND ADDRESS Dept. of Psychological Sciences Purdue University West Lafayette, IN 47907		8. CONTRACT OR GRANT NUMBER(s) N00014-82-K-0449
11. CONTROLLING OFFICE NAME AND ADDRESS Organizational Effectiveness Research Programs Office of Naval Research, Arlington, VA 22217		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61150N 42 NR 170-91 TR04138 RR0420001
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		12. REPORT DATE March, 1983
		13. NUMBER OF PAGES 29
		15. SECURITY CLASS. (of this report) Unclassified
		19a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the U.S. Government.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES B		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Mixed Standard Scale, Performance Ratings, Extremity		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) It was suggested that the extremity of the scale values associated with standards used to represent effective and ineffective performance in Mixed Standard Scales may affect the nature of performance ratings derived from MSS responses and decisions based on MSS ratings. When the extremity of standards was experimentally manipulated, it was found that standard extremity affects both the level of performance ratings and the proportion of logically inconsistent response patterns observed. In addition, standard		

DTIC  
ELECTE  
MAY 27 1983

S D B

AD A128645

DTIC FILE COPY

DD FORM 1473  
1 JAN 73EDITION OF 1 NOV 65 IS OBSOLETE  
S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

extremity appears to affect the rankings on performance of ratees. The implications of these observations for the development of Mixed Standard Scales were discussed.

Effects of Standard Extremity on Mixed Standard Scale Performance Ratings

Janet L. Barnes-Farrell and Howard M. Weiss

Purdue University

Prepared for  
Organizational Effectiveness Research Programs  
Office of Naval Research

Contract N00014-82-K-0449  
NR 170-940

Approved for public release; distribution unlimited.  
Reproduction in whole or in part is permitted for  
any purpose of the United States Government.

83 05 27 037

# Abstract

It was suggested that the extremity of the scale values associated with standards used to represent effective and ineffective performance in Mixed Standard Scales may affect the nature of performance ratings derived from MSS responses and decisions based on MSS ratings. When the extremity of standards was experimentally manipulated, it was found that standard extremity affects both the level of performance ratings and the proportion of logically inconsistent response patterns observed. In addition, standard extremity appears to affect the rankings on performance of ratees. The implications of these observations for the development of Mixed Standard Scales were discussed.



Accession For	
NTIS GRA&I	<input checked="checked" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A	

## Effects of Standard Extremity on Mixed Standard Scale Performance Ratings

In 1972, Blanz and Ghiselli introduced the Mixed Standard Scale (MSS) approach to rating employee performance. Like the more popular BARS approach, the MSS procedure assumes that raters will make more accurate and reliable judgments about the levels at which their employees are performing if they are provided with descriptions of the kinds of behaviors characterizing effective and ineffective performance on each performance dimension. Unlike BARS, the MSS is a derived scale, in which neither the performance dimension nor the effectiveness level of anchor statements is provided to raters when they use the scales. Rather than being asked to compare each ratee to a continuum of performance effectiveness for each dimension, the rater is asked to compare the ratee's performance to a series of statements (standards) representing varying levels of performance effectiveness and varying performance dimensions. The standards are "mixed" (presented in a random order) so that neither the effectiveness levels nor the performance dimensions they represent are readily apparent to the rater. For each statement, the rater must decide whether ratee performance equals, surpasses, or is less effective than the performance level exemplified in the standard. The patterns of responses to the standards representing each performance dimension are then transformed into dimension ratings on a 7-point scale.

Because the underlying rating scale is disguised to the rater, Blanz and Ghiselli expected that such rater biases as leniency and halo would be reduced. In addition, since mixed standard scales are assumed to have Guttman properties, the patterns of responses that raters exhibit can be indexed in terms of their logical consistency<sup>2</sup>. Raters with high levels of logical inconsistency can be identified, and perhaps be given special attention or training. Likewise, ratees for whom high levels of logical error are observed can be identified.

Despite the stated advantages of the format, the MSS approach to rating employee performance has received only intermittent attention from industrial psychologists in the last ten years. Most examinations of the MSS format have focused on one of three issues:

- a) difficulties associated with deriving a consistent coding system for transforming item responses into dimension ratings (Saal, 1979);
- b) the effect of anchor content and developmental procedures on the psychometric characteristics of ratings obtained with MSS (Dickinson & Zellinger, 1980);
- c) comparisons of the psychometric characteristics of ratings obtained from MSS and other rating formats (Arvey & Hoyle, 1974; Dickinson & Zellinger, 1980; Finley, Osburn, Dublin, & Jeanneret, 1977; Saal & Landy, 1977; Saal, 1979)

In general, evaluations of the MSS format have been mixed. While most examinations of leniency have concluded that the



mixed standard scale format performs at least as well as, and sometimes better than, the BARS format or simple graphic rating scale (Finley, et al., 1977; Saal, 1979; Saal & Landy, 1977), conclusions regarding the relative effectiveness of the MSS format in reducing levels of halo have been inconsistent, sometimes favoring MSS (Saal, 1979; Saal & Landy, 1977), and sometimes favoring BARS (Arvey & Hoyle, 1974; Finley, et al., 1977). Lack of inter-rater reliability does seem to be a consistent problem with the MSS (Arvey & Hoyle, 1974; Finley, et al., 1977; Saal, 1979; Saal & Landy, 1977). However, the convergent and discriminant validity of ratings obtained with a mixed standard format appears to be acceptable and equivalent to that observed in ratings obtained with a BARS format, as long as similar developmental procedures (i.e. behavioral anchors and retranslation of expectations) are used to produce the scales (Arvey & Hoyle, 1974; Dickinson & Zellinger, 1980).

Since the number of behavioral examples anchoring each performance dimension is very small, and raters are provided no information about the relative or absolute performance levels that these anchors are intended to represent, the nature and underlying scale values of the anchors chosen to describe a performance dimension have potentially important implications for the ways in which raters respond to the instrument, and the ratings that are derived from those responses. Yet little is known about the manner in which anchors are chosen for mixed standard scales, or the influence that this aspect of the development process might have on performance descriptions.

The current study addresses this issue. Specifically, we were interested in the impact of anchor selection procedures on the ratings obtained with a rating instrument that utilizes a mixed standard format.

Consider the typical recommended procedure for constructing a mixed standard scale:

Step 1: Generate and define performance dimensions to be evaluated.

Step 2: Generate critical incidents describing different levels of performance for each dimension (anchors, or "standards").

Step 3: For each performance dimension, choose a standard representing high, moderate, and low levels of effectiveness, respectively.

Step 4: Prepare a final list of performance standards which has been mixed across performance dimensions and across performance levels (i.e., if there are X dimensions, the final list will have 3X performance standards to which the rater must respond).

Our questions grew out of a consideration of Step 3. Since ratings obtained with a Mixed Standard Scale are derived scores, the scale values of the standards to which raters respond are ignored once the standard has been assigned to the category high (H), moderate (M), or low (L). Instead, dimension ratings are based on the rater's patterns of responses to the chosen standards for each dimension. However, no standard decision rules have been presented to guide the

Instrument developer In selecting standards to represent the various performance levels (with the exception, of course, that the resulting scales should have Guttman properties). For instance, if we think of the various standards as representing various levels on a seven-point scale, the developer might choose standards with scale values of 6, 4, and 2 to represent the categories H, M, and L respectively. Alternatively, he/she might choose standards with scale values of 7, 4, and 1 to represent the same categories. The mixed standard scales produced by these decision rules vary in terms of: (1) the extremity of scale values underlying each rating dimension; and (2) the amount of scale separation among standards representing different levels of performance. (The two are of course not independent of one another, since the extremity of the scale values constrains the amount of scale separation among standards.) These variations may affect rater responses to the scales, with implications for the psychometric characteristics of the ratings obtained, and for decisions which are based upon those ratings.

First, consider the way in which extremity of chosen standards (hereafter referred to as "standard extremity") might influence the level of ratings assigned with a MSS. Raters are asked to decide whether each ratee performs at (0), above (+), or below (-) the level of each standard presented to them. The pattern of three responses to each dimension is transformed into a rating on a seven-point scale. The probability of responding +, -, or 0 to a particular standard should be

affected by the performance level of the ratee. However, it will also be constrained by the extremity of the anchors chosen to represent high and low performance. Behaviors at the extreme ends of the scale will be relatively rare. Raters who are responding to standards chosen from the extreme ends would thus be less likely to have observed behaviors at those levels than would raters responding to less extreme standards. As such, we would expect that raters using a mixed standard scale comprised of high performance standards (H) with scale values of 7 would be less likely to respond with the pattern of responses which is transformed to a rating of 7 (+++) than would raters using a scale with high performance standards having a scale value of 6. A similar situation should occur when raters attempt to provide ratings of low performance levels. This would result in decreased variability in assigned ratings but no change in the level of ratings if performance is normally distributed; but generally this is not the case. Typically, actual performance distributions in organizations are negatively skewed. When the distribution is skewed, we would expect standard extremity to have a linear effect on the level of ratings assigned. The implication for the distribution of ratings derived from rater response patterns will be increased central tendency in ratings gathered from mixed standard scales whose performance standards have extreme scale values.

The amount of scale separation among performance standards, on the other hand, would be expected to affect the degree to which raters are able to reliably differentiate among

performance levels. Performance standards representing scale values of 1, 4, and 7 should be more readily distinguished and rank-ordered than performance standards with scale values of 2, 4, and 6, for example. The latter are perceptually more similar to one another in terms of performance level. As a result, we might expect to see an increase in the frequency of logical errors present in ratings as the distance between anchor statements decreases.

The current study tested both of these hypotheses at the effect of developmental procedures on the characteristics of ratings obtained when a MSS is used to evaluate performance. In addition, two other issues were examined. Since one of the intended advantages of the "mixed" format of the MSS is the reduction of halo, it is reasonable to ask whether anchor extremity (and resulting decreased anchor separation) affects halo. Finally, we thought it important to consider the practical implications that anchor selection procedures might have for decisions made on the basis of inter-individual comparisons. For example, when a promotion decision is being made by a supervisor or personnel department, most often the task is one of rank-ordering eligible employees in terms of some criterion of performance effectiveness or potential to perform. When we use a mixed standard scale to differentiate among employees in this way, does the extremity of the performance standards in the MSS affect the rank-ordering of ratees, and ultimately the decisions of the organization?

## METHOD

Subjects. Subjects were 248 students recruited from the classes of seven Introductory psychology instructors who agreed to participate in this study. Participation in the study was voluntary.

Materials. Three mixed standard scales for the evaluation of teacher performance were prepared from a pool of statements previously developed for use in behavioral expectation scales by Harari and Zedeck (1973). These materials were chosen specifically because they represented an example of performance appraisal scales that met several important criteria:

- 1) Behaviorally anchored - the content of the anchors was behavioral and specific;
- 2) Rigorous development procedures-the Harari and Zedeck scales were carefully developed using the retranslation of expectations (RE) technique to eliminate anchors which were not unambiguous examples of performance dimensions, and a second screening to eliminate those anchors for which there was disagreement about the effectiveness level (scale value) represented.
3. Multiple anchor points - the behavioral anchors represented a range of scale values which could be easily translated into mixed standard scales having the variations in standard extremity that we required to test our hypothesis;

4. Invariance of scale values - because the behavioral anchors used in this study were developed and scaled in another setting, there was some concern that the scale values might not generalize to settings other than the one in which the scales were developed. However, a study by Landy and Barnes (1979) which used statements from the same pool, indicated that the mean scale values assigned to the behavioral statements developed by Harari and Zedeck (1973) did not change when those statements were rescaled several years later at a second university.

Each MSS was composed of a total of twelve statements representing high, moderate, and low levels of performance effectiveness in four areas: Delivery, Ability to Motivate Student, Depth of Knowledge, and Interpersonal Relations with Students. Information about the scale values of behavioral statements defining each dimension was used to construct three different mixed standard scales, varying in terms of the extremity of the scale values associated with the standards defining high and low effectiveness levels. MSS I (HE) was composed of statements reflecting maximally extreme scale values for each of the four dimensions represented on the appraisal instrument. MSS II (ME) was composed of statements with moderately extreme scale values. MSS III (LE) was composed of statements with minimally extreme scale values. The scale values associated with the standards comprising each of the three mixed standard scales are presented in Table 1.

-----  
Insert Table 1 about here  
-----

Procedure. Students of the seven introductory psychology instructors who agreed to participate in this study were randomly assigned to one of three conditions: High Extremity (HE), Moderate Extremity (ME), or Low Extremity (LE). All subjects were told that they would be evaluating the performance of their instructor, and that their evaluations would be used to provide feedback to their instructor about his/her performance strengths and weaknesses. Each subject rated only one instructor. Each instructor was evaluated by 23 to 53 students.

Analyses. Responses to the mixed standard scales were coded to produce performance dimension ratings on a 7-point scale, using the coding scheme suggested by Saal (1979).

Means and standard deviations for assigned dimension ratings were calculated for each of the experimental conditions (HE, ME, LE) and for each rater. In addition, dimension intercorrelation matrices were constructed for each of the three experimental conditions. Finally, a simple tally of the number of inconsistency errors (response patterns inconsistent with the scaled order of standards for each dimension) was computed for each experimental condition.

To test the hypothesis that standard extremity affects the level of central tendency, central tendency was operationalized as a level effect. A multivariate analysis of variance (MANOVA) linear trend analysis was performed to test the effect of standard extremity on performance ratings. This was



followed by one-way analyses of variance (ANOVAs) for each performance dimension, using standard extremity as the independent variable, and assigned performance rating as the dependent variable.

To test the hypothesis that the amount of "logical" error is affected by standard extremity, all dimension ratings were scored as consistent (no logical error=0) or inconsistent (logical error=1). That is, if the set of responses to a dimension was one of the 7 logically consistent response combinations, the rating derived from that set of responses was said to be logically consistent. A rating derived from any one of the 20 logically inconsistent response combinations was said to be logically inconsistent. Although there are many ways to provide consistent ratings (7 ways) and inconsistent ratings (20 ways), each rater could only commit one logical error per performance dimension. A one-way ANOVA using standard extremity as the independent variable and proportion of logical errors in the observed ratings as the dependent variable, was performed for each performance dimension after a MANOVA for linear trends was used to test the effect of standard extremity on logical errors.

In order to examine the question of whether halo is affected by standard extremity, halo was operationalized in two ways. The first index of halo was defined as the mean intercorrelation between dimension ratings assigned in each condition. To compute mean intercorrelation levels, a Fisher Z transformation was applied to the zero-order intercorrelation

matrices. A chi-square test for homogeneity was used to test the hypothesis that levels of halo are different for different experimental conditions. Halo was also operationalized as the standard deviation of each rater's ratings across the four performance dimensions (where high standard deviations indicate low halo levels). In order to use standard deviations as data points, a log transformation was applied. A one-way ANOVA was performed, using standard extremity as the independent variable.

Finally, the practical implications of variations in standard extremity were examined by rank-ordering instructors on the basis of the mean performance ratings assigned to them for each dimension and the overall mean ratings assigned to them. A rank order correlation between HE ratings and ME ratings was computed for each dimension and for the overall mean summated ratings. The same comparison was made between ME ratings and LE ratings, and between HE ratings and LE ratings. Since the number of teachers being ranked was small ( $n = 7$ ) tau rather than Spearman's rho was used (Thorndike, 1978). However, tau ranges from -1.0 to +1.0 and is interpreted in the same manner as rho.

## RESULTS

A MANOVA using performance ratings on all four performance dimensions as dependent variables and standard extremity as an independent variable indicated that standard extremity significantly affects the level of assigned ratings. Further, the analysis indicated a significant linear trend (F

approximation for Pillai-Bartlett  $V=4.54$ ;  $df=4,240$ ;  $p<.01$ ). The results of followup univariate ANOVAs conducted separately for each performance dimension can be seen in Table 2.

-----  
Insert Table 2 about here  
-----

Standard extremity had a significant effect ( $p<.001$ ) on the level of performance ratings assigned for two performance dimensions: Ability to Motivate and Depth of Knowledge, and a marginally significant effect for the remaining two dimensions: Delivery ( $p<.09$ ) and Interpersonal Relations with Students ( $p<.06$ ). An examination of the cell means for each dimension (also shown in Table 2) indicates a pattern of results generally consistent with our hypothesis that central tendency will increase as the extremity of scale values underlying standards of high and low performance increases. For all four dimensions, mean ratings were closest to the center of the scale for the high extremity condition. Post hoc linear trend analyses showed significant linear trends in the data for the first three dimensions ( $p < .05$ ) and a marginally significant linear component for the fourth dimension ( $p < .08$ ).

In addition to the analysis for the total sample, a similar analysis was performed on that subset of rater responses consisting only of those ratings representing logical response patterns. This was done in order to explore what effects standard extremity might have when ratings are uncontaminated by the error variance introduced when raters respond in logically inconsistent ways. In other words, we

were interested in identifying whether a level effect would still be observed for those cases in which the MSS was used as it was intended to be used, free from logical inconsistency errors. We found that this secondary analysis makes the pattern of results even clearer (see Table 3). For all three dimensions in which a significant main effect was observed

-----  
Insert Table 3 about here  
-----

(Delivery, Ability to Motivate and Depth of Knowledge) the means were ordered in the expected pattern, and significant linear trends were found ( $p < .01$ ).

A MANOVA using standard extremity as the independent variable and proportions of logical inconsistency error in each of the four performance dimensions as dependent variables also supported the hypothesis that the amount of logical error present in MSS ratings is affected by standard extremity. As with the previous analyses the effect of the independent variable had a significant linear component (F approximation for Pillai-Bartlett  $V = 3.68$ ;  $df = 4, 242$ ;  $p < .01$ ). However as can be seen in Table 4, univariate ANOVAs conducted for each of

-----  
Insert Table 4 about here  
-----

the four performance dimensions indicated a significant main effect for only one dimension: Ability to Motivate ( $p < .001$ ). The cell means for Ability to Motivate show decreasing levels of logical inconsistency error as standard extremity increases, as predicted (linear trend,  $F = 13.38$ ,  $df = 1, 245$ ,  $p < .01$ ).

Halo was not affected by standard extremity. The mean

Intercorrelation between dimension ratings ranged from  $r=.34$  to  $r=.38$  ( $\chi^2 = .10$ ,  $df=2$ , n.s.) for the three experimental conditions. The standard deviation of each rater's assigned ratings across the four dimensions ranged from 1.39 to 1.57 ( $F=.46$ ;  $df=2,243$ ; n.s.).

Finally, examination of the rank-ordering of Instructors which is produced by performance ratings provides evidence that the extremity of scale anchors in a MSS affects the rank order of Instructors. Values of tau summarizing the similarity of rank-orderings produced under different experimental conditions are reported in Table 5. Examination of this table reveals that the magnitude of the tau statistic for the rank-order comparisons was low. Only 3 of the 15 tau coefficients

-----  
Insert Table 5 about here  
-----

calculated were significantly greater than 0. None of the comparisons between HE and LE conditions or between ME and LE conditions produced significant rank-order associations. The only significant correlations were observed in rank orderings produced in the HE and ME conditions which were similar for Delivery, Ability to Motivate Students and Overall Mean Rating ( $\tau = .71, .90$ , and  $.71$  respectively).

#### DISCUSSION

The results of this study generally supported our hypotheses that the extremity of the scale values associated with standards chosen in the development of mixed standard scales affects 1) the level of ratings assigned, and 2) the

number of logically inconsistent response patterns which are exhibited; and 3) the relative position of respondents in performance distributions.

Support for the first hypothesis was relatively consistent, indicating a tendency for ratings to be assigned closer to the center of the scale as standard extremity increased. This effect became more pronounced when we examined the subset of ratings which conformed to one of the seven logically consistent response patterns. Presumably, these ratings are free of some of the "noise" contaminating the full set of ratings. Yet it is apparent that the noise introduced when raters respond to mixed standard scales in logically inconsistent ways only masks the underlying phenomenon to some extent. Thus, attempts to improve the quality of ratings by training raters to respond carefully (in logically consistent patterns) will only make developmental issues like this one more important. From a practical standpoint, the organization in the process of developing or revising a performance appraisal instrument using a mixed standard format, can use this information to advantage. For example, if positive leniency is a problem, careful attention should be paid to choosing high and low effectiveness standards with scale values falling as close to the extreme ends of the scale as possible. In any case, the instrument developer should be aware of the fact that all examples of highly effective (or highly ineffective) performance are not necessarily equivalent, and that the choice of standards that is made may affect the level

of ratings assigned. This issue might be of particular importance for performance appraisal systems in which an attempt is made to measure several dimensions of employee performance and then form a profile of employee strengths and weaknesses. If attention is not paid to the issue of underlying scale values, the rank ordering of an employee's weaknesses might reflect a rank ordering of performance dimensions on the basis of standard extremity rather than a measure of employee performance on dimension A relative to dimension B, etc. As we pointed out in the Introduction, standard extremity would only be expected to affect the level of assigned ratings when the distribution of actual performance is skewed. Although we have no way of determining whether this was the case in the sample of ratees that we observed, there is good reason to believe that negatively skewed performance distributions are typical in organizations (cf. Bernardin & Pence, 1980) and in samples of teachers in particular (Zedeck, Jacob, & Kafry, 1976).

Support for our hypothesis regarding the effect of standard extremity on the proportion of logically inconsistent response patterns observed was weaker. The expected increase in logical inconsistency errors as the scale separation of standards decreases was only observed for one dimension: Ability to Motivate Students. While the choice of standard didn't influence the error rate as expected, it is significant to note the high frequency of these "error" responses. Even when using scales that have been carefully developed using retranslation of expectations procedures to ensure that

standards unambiguously represent performance dimensions, and response scaling to ensure that raters agree on the effectiveness level represented by each anchor, approximately half of the performance ratings collected were derived from patterns of responses that were, in one way or another, logically inconsistent. The high frequency of logical errors may be, in part, a reflection of the motivation of student raters to do a careful job in evaluating their instructors' performance and recording those evaluations. (It was for this reason that we felt that the secondary analysis of the rating level effect data was necessary and useful.) On the other hand, low motivation may be typical in many organizational settings. Since little normative data on the frequency of inconsistent response patterns is available in the published literature, it is difficult to say whether our data were contaminated by an unusually large proportion of illogical responses or whether the "error" rate in our data is similar to that obtained in other studies.

Because each rater only evaluated the performance of one instructor, it is difficult to make assessments about the degree to which logical errors were primarily a rater effect rather than an instrument effect or a ratee effect. However, an examination of the intercorrelation matrix summarizing the relationships between error scores on different dimensions (i.e., error or no error, since a rater can only make one error per dimension) revealed significant but small correlations



(mean correlation between dimensions  $\phi = +.14$ ,  $\chi^2 = 24.6$ ,  $p < .001$ ). That is, raters who respond with inconsistent patterns on one dimension are slightly more likely to make logical errors on other dimensions. Still, the very small proportion of raters who provided a complete set of responses free from logical inconsistency errors (only 9% of the sample: 21 of 248 raters) suggests that inconsistency errors are a rather general feature of this set of ratings, rather than a problem limited to a small number of raters. To examine the possibility of a ratee effect on logical inconsistency errors, an eta coefficient between ratees and the total number of logical errors per rater was calculated. Eta-squared was only .01 (n.s.), suggesting very little (if any) relationship between teachers and the tendency to make logical errors in evaluating them.

It seems reasonable to conclude, then, that the problem of logical inconsistency errors is not one which can be primarily attributed to individual differences in raters or ratees, but is more likely associated with the instrument and the way that raters respond to a mixed standard scale format. The magnitude of the problem is such that research directed at understanding the conditions which influence the manner in which raters respond to performance appraisal instruments with disguised continua is necessary if we are to have any confidence in the ratings derived from such scales. If the major source of variance is motivational, it might be reasonable to suggest training or some similar intervention as a strategy for decreasing the problem. On the other hand, if the source of

the problem is related to cognitive strategies which are typically used by raters in processing and evaluating performance information, we may find that the wiser course would be to modify the process of recording performance evaluations so that they are more compatible with rater cognitive strategies.

The observation that standard extremity may affect the rank-ordering of ratees, both for individual performance dimensions and for the overall mean of dimension ratings clearly indicates that developmental procedures will affect personnel decisions based on performance ratings. The effect of standard extremity on interindividual comparisons is not a simple one, and we can offer no straightforward explanation for why rank orderings change in the way that they do. We also have no reason to believe that one rank-ordering is more accurate than another. However, the mere fact that rank orders change as a function of anchor selection procedures suggests that organizations need to pay close attention to the underlying scale values of standards chosen to represent effective and ineffective performance when developing mixed standard scales. This is particularly important when several different forms will be developed (e.g., rating scales tailored to particular groups of job titles), and when important decisions (e.g. selection for promotion) will be based upon a rank-ordering of employees according to the performance ratings assigned to them.

#### Footnotes

1. This research was supported in part by U.S. Office of Naval Research Contract N00014-82-K-0449 (Janet L. Barnes-Farrell and Daniel R. Ilgen, co-principle investigators); and in part by U.S. Office of Naval Research Contract N00014-78-C-0609 (Howard M. Weiss, principle investigator).

The authors wish to thank Sheldon Zedeck, who graciously supplied the scales used in the Harari and Zedeck (1973) study, as well as the means scale values for the anchors. Requests for reprints should be sent to Janet L. Barnes-Farrell, Department of Psychological Sciences, Purdue University, West Lafayette Indiana 47907.

2. By logical consistency, we mean that responses conform to Guttman scale assumptions. Any set of responses to items on a Guttman scale which forms a pattern that does not conform with those assumptions is said to be logically inconsistent. In the context of mixed standard scales, each rating is derived from the pattern of responses (+, 0, or -) to three statements representing different levels of performance effectiveness. There are 27 possible response combinations to each set of three statements. Seven of these response combinations are logically consistent with the patterns of responses that would be expected if those statements formed a Guttman scale: they are said to be logically consistent. The remaining 20 response combinations are not consistent with the patterns

of responses that would be expected if those statements formed a Guttman scale. When any of those response combinations is observed, it is referred to as a logical inconsistency error.

## References

- Arvey, R. A. & Hoyle, J. C. A Guttman approach to the development of behaviorally based rating scales for systems analysts and programmer/analysts. Journal of Applied Psychology, 1974, 59, 61-68.
- Bernardin, H. J. & Pence, E. C. Effects of rater training: Creating new response sets and decreasing accuracy. Journal of Applied Psychology, 1980, 65, 60-66.
- Blanz, F. & Ghiselli, E. The mixed standard scale: A new rating system. Personnel Psychology, 1972, 25, 185-199.
- Dickinson, T. L. & Zellinger, P. M. A comparison of behaviorally anchored rating and mixed standard scale formats. Journal of Applied Psychology, 1980, 65, 147-154.
- Finley, D., Osburn, H., Dubin, J. & Jeanneret, P. Behaviorally based rating scales: Effects of specific anchors and disguised scale continua. Personnel Psychology, 1977, 30, 659-669.
- Harari, O. & Zedeck, S. Development of behaviorally anchored scales for the evaluation of faculty teaching. Journal of Applied Psychology, 1973, 58, 261-265.
- Landy, F. & Barnes, J. Scaling behavioral anchors. Applied Psychological Measurement, 1979, 3, 193-200.
- Saai, F. E. Mixed standard rating scale: A consistent system for numerically coding inconsistent response combinations. Journal of Applied Psychology, 1979, 64, 422-428.

Saal, F. & Landy, F. The mixed standard rating scale: An evaluation. Organizational Behavior and Human Performance, 1977, 18, 19-35.

Thorndike, R. M. Correlational procedures for research. New York: Gardner Press, 1978.

Zedeck, S., Jacob, R., & Kafry, D. Behavioral expectations: Development of parallel forms and analysis of scale assumptions. Journal of Applied Psychology, 1976, 61, 112-115.

Table 1

## Scale Values for Standards Used to Create Mixed Standard Scales

Performance Effectiveness Level		Performance Dimension			
		Delivery	Ability to Motivate	Knowledge	Interpersonal Relations
High:	HE	6.4	6.4	6.4	6.5
	ME	5.9	6.0	5.7	5.8
	LE	5.0	5.0	4.2	4.9
Moderate: All conditions		3.8	3.8	3.8	3.9
Low:	LE	2.7	2.4	3.0	2.9
	ME	2.3	2.0	2.1	2.1
	HE	1.5	1.6	1.4	1.3

a

HE = High Extremity condition

ME = Moderate Extremity condition

LE = Low Extremity condition

Note: For examples of the kinds of statements used as standards for each dimension, see Harari & Zedeck (1973) or Landy & Barnes (1979).

Table 2

Cell Means and Univariate F-tests for the Effects of Standard Extremity  
on Performance Ratings

<u>Performance Dimension</u>	<u>ANOVA Summary Table</u>						<u>Cell Means</u> <sup>a</sup>		
	Source	SS	df	MS	F	p-level	HE	ME	LE
Delivery	Extremity	11.02	2	5.51	2.42	<.09	4.11	4.36	4.63
	Residual	554.05	243	2.28					
	Total	565.07	245	2.31					
Ability to Motivate	Extremity	62.82	2	31.41	20.68	<.001	3.88	5.10	4.29
	Residual	369.05	243	1.52					
	Total	431.87	245	1.76					
Depth of Knowledge	Extremity	39.02	2	19.51	8.91	<.001	4.57	4.99	5.54
	Residual	531.65	243	2.19					
	Total	570.67	245	2.33					
Inter- personal Relations	Extremity	10.98	2	5.47	2.91	<.05	4.84	5.33	5.23
	Residual	457.59	243	1.88					
	Total	468.57	245	1.91					

a

HE = High Extremity condition  
ME = Moderate Extremity condition  
LE = Low Extremity condition



Table 3

Cell Means and Univariate F-tests for the Effects of Standard  
Extremity on Performance Ratings: Logical Responses Only

<u>Performance Dimension</u>	<u>ANOVA Summary Table</u>						<u>Cell Means</u>		
	Source	SS	df	MS	F	p-level	HE	ME	LE
Delivery (N = 105)	Extremity	33.12	2	16.56	5.32	<.01	4.09	5.06	5.44
	Residual	317.51	102	3.11					
	Total	350.63	104	3.37					
Ability to Motivate (N = 94)	Extremity	35.87	2	17.94	10.14	<.001	4.27	5.45	5.60
	Residual	161.03	91	1.77					
	Total	196.90	93	2.12					
Depth of Knowledge (N = 116)	Extremity	39.22	2	19.61	9.66	<.001	5.05	5.97	6.42
	Residual	229.33	113	2.03					
	Total	268.55	115	2.34					
Inter- personal Relations (N = 101)	Extremity	2.94	2	1.47	.95	>.10	5.97	6.18	6.11
	Residual	151.98	98	1.55					
	Total	154.91	100	1.55					

Table 4

Cell Means and Univariate F-tests for the Effects of Standard Extremity  
on Proportion of Logical Errors

<u>Performance Dimension</u>	<u>ANOVA Summary Table</u>						<u>Cell Means</u>		
	Source	SS	df	MS	F	p-level	HE	ME	LE
Delivery	Extremity	.08	2	.04	.17	>.10	.60	.55	.57
	Residual	60.61	245	.25					
	Total	60.69	247	.25					
Ability to Motivate	Extremity	3.06	2	1.53	6.75	<.001	.49	.60	.76
	Residual	55.55	245	.23					
	Total	58.61	247	.24					
Depth of Knowledge	Extremity	.50	2	.25	.99	>.10	.52	.59	.48
	Residual	61.25	245	.25					
	Total	61.74	247	.25					
Inter- personal Relations	Extremity	.02	2	.01	.03	>.10	.60	.59	.58
	Residual	60.03	245	.25					
	Total	60.04	247	.24					

Table 5  
 Association (M) between Rank-orderings Produced  
 by Different Experimental Conditions

<u>Performance Dimension</u>	<u>Rank-order Correlation</u>		
	HE and ME	ME and LE	HE and LE
Delivery	.71*	.24	.14
Ability to Motivate	.90**	.43	.33
Depth of Knowledge	-.11	-.24	.43
Interpersonal Relations	.33	.24	.33
Mean Overall Rating	.71*	.30	.40

\*p < .05

\*\*p < .01

Copy available to DTIC does not  
 permit fully legible reproduction

LIST 1  
MANDATORY

Defense Technical Information Center  
ATTN: DTIC DDA-2  
Selection and Preliminary Cataloging Section  
Cameron Station  
Alexandria, VA 22314

Library of Congress  
Science and Technology Division  
Washington, D.C. 20540

Office of Naval Research  
Code 4420E  
800 N. Quincy Street  
Arlington, VA 22217

Naval Research Laboratory  
Code 2627  
Washington, D.C. 20375

Office of Naval Research  
Director, Technology Programs  
Code 200  
800 N. Quincy Street  
Arlington, VA 22217

LIST 2  
ONR Field

Psychologist  
Office of Naval Research  
Detachment, Pasadena  
1030 East Green Street  
Pasadena, CA 91106

Dr. James Lester  
Office of Naval Research  
Detachment, Boston  
495 Summer Street  
Boston, MA 02219

LIST 3  
OPNAV

Deputy Chief of Naval Operations  
(Manpower, Personnel, and Training)  
Head, Research, Development, and  
Studies Branch (Op-115)  
1812 Arlington Annex  
Washington, D.C. 20350

Director  
Civilian Personnel Division (OP-14)  
Department of the Navy  
1803 Arlington Annex  
Washington, D.C. 20350

Deputy Chief of Naval Operations  
(Manpower, Personnel, and Training)  
Director, Human Resource Management  
Plans and Policy Branch (Op-150)  
Department of the Navy  
Washington, D.C. 20350

Chief of Naval Operations  
Head, Manpower, Personnel,  
Training and Reserves Team  
(Op-964D)  
The Pentagon, 4A478  
Washington, D.C. 20350

Chief of Naval Operations  
Assistant, Personnel Logistics  
Planning (Op-987H)  
The Pentagon, 5D772  
Washington, D.C. 20350

LIST 4  
NAVMAT & NRPDC

NAVMAT

Program Administrator for Manpower,  
Personnel, and Training

MAT-0722  
800 N. Quincy Street  
Arlington, VA 22217

Naval Material Command  
Management Training Center  
NAVMAT 09M32  
Jefferson Plaza, Bldg #2, Rm 150  
1421 Jefferson Davis Highway  
Arlington, VA 20360

Naval Material Command  
MAT-00K & MAT-00KB  
OASN(SNL)  
Crystal Plaza #5  
Room 236  
Washington, D.C. 20360

Naval Material Command  
MAT-03  
(J. E. Colvard)  
Crystal Plaza #5  
Room 236  
Washington, D.C. 20360

NRPDC

Commanding Officer  
Naval Personnel R&D Center  
San Diego, CA 92152

Naval Personnel R&D Center  
Dr. Robert Penn  
San Diego, CA 92152

Naval Personnel R&D Center  
Dr. Ed Aiken  
San Diego, CA 92152

Navy Personnel R&D Center  
Washington Liaison Office  
Building 200, 2N  
Washington Navy Yard  
Washington, D.C. 20374

LIST 6  
NAVAL ACADEMY AND NAVAL POSTGRADUATE SCHOOL

Naval Postgraduate School  
ATTN: Dr. Richard S. Elster (Code 012)  
Department of Administrative Sciences  
Monterey, CA 93940

Naval Postgraduate School  
ATTN: Professor John Senger  
Operations Research and  
Administrative Science

Superintendent  
Naval Postgraduate School  
Code 1424  
Monterey, CA 93940

Naval Postgraduate School  
Code 54-Aa  
Monterey, CA 93940

Naval Postgraduate School  
ATTN: Dr. Richard A. McGonigal  
Code 54  
Monterey, CA 93940

U.S. Naval Academy  
ATTN: CDR J. M. McGrath  
Department of Leadership and Law  
Annapolis, MD 21402

Professor Carson K. Eoyang  
Naval Postgraduate School, Code 54EG  
Department of Administrative Sciences  
Monterey, CA 93940

Superintendent  
ATTN: Director of Research  
Naval Academy, U.S.  
Annapolis, MD 21402

# LIST 7

## HRM

Officer in Charge  
Human Resource Management Detachment  
Naval Air Station  
Alameda, CA 94591

Officer in Charge  
Human Resource Management Detachment  
Naval Submarine Base New London  
P. O. Box 81  
Groton, CT 06340

Officer in Charge  
Human Resource Management Division  
Naval Air Station  
Mayport, FL 32228

Commanding Officer  
Human Resource Management Center  
Pearl Harbor, HI 96860

Commander in Chief  
Human Resource Management Division  
U.S. Pacific Fleet  
Pearl Harbor, HI 96860

Officer in Charge  
Human Resource Management Detachment  
Naval Base  
Charleston, SC 29408

Commanding Officer  
Human Resource Management School  
Naval Air Station Memphis  
Millington, TN 38054

Human Resource Management School  
Naval Air Station Memphis (96)  
Millington, TN 38054

Commanding Officer  
Human Resource Management Center  
1300 Wilson Boulevard  
Arlington, VA 22209

Commanding Officer  
Human Resource Management Center  
5621-23 Tidewater Drive  
Norfolk, VA 23511

Commander in Chief  
Human Resource Management Division  
U.S. Atlantic Fleet  
Norfolk, VA 23511

Officer in Charge  
Human Resource Management Detachment  
Naval Air Station Whidbey Island  
Oak Harbor, WA 98278

Commanding Officer  
Human Resource Management Center  
Box 23  
FPO New York 09510

Commander in Chief  
Human Resource Management Division  
U.S. Naval Force Europe  
FPO New York 09510

Officer in Charge  
Human Resource Management Detachment  
Box 60  
FPO San Francisco 96651

Officer in Charge  
Human Resource Management Detachment  
COMNAVFORJAPAN  
FPO Seattle 98762

# LIST 8

## NAVY MISCELLANEOUS

Naval Military Personnel Command  
HRM Department (NMPC-6)  
Washington, D.C. 20350

LIST 15  
CURRENT CONTRACTORS

Dr. Clayton F. Alderfer  
Yale University  
School of Organization and Management  
New Haven, Connecticut 06520

Dr. Richard D. Arvey  
University of Houston  
Department of Psychology  
Houston, TX 77004

Dr. Stuart W. Cook  
Institute of Behavioral Science #6  
University of Colorado  
Box 482  
Boulder, CO 80309

Dr. L. L. Cummings  
Kellogg Graduate School of Management  
Northwestern University  
Nathaniel Leverone Hall  
Evanston, IL 60201

Dr. Richard Daft  
Texas A&M University  
Department of Management  
College Station, TX 77843

Bruce J. Bueno De Mesquita  
University of Rochester  
Department of Political Science  
Rochester, NY 14627

Dr. Henry Emurian  
The Johns Hopkins University  
School of Medicine  
Department of Psychiatry and  
Behavioral Science  
Baltimore, MD 21205

Dr. Arthur Gerstenfeld  
University Faculty Associates  
710 Commonwealth Avenue  
Newton, MA 02159

Dr. Paul S. Goodman  
Graduate School of Industrial  
Administration  
Carnegie-Mellon University  
Pittsburgh, PA 15213

Dr. J. Richard Hackman  
School of Organization  
and Management  
Box 1A, Yale University  
New Haven, CT 06520

Dr. Herry Hunt  
College of Business Administration  
Texas Tech. University (Box 4320)  
Lubbock, TX 79409

Dr. Lawrence R. James  
School of Psychology  
Georgia Institute of Technology  
Atlanta, GA 30332

Dr. F. Craig Johnson  
Department of Educational Research  
Florida State University  
Tallahassee, FL 32306

Dr. Allan P. Jones  
University of Houston  
4800 Calhoun  
Houston, TX 77004

Dr. Dan Landis  
Department of Psychology  
Purdue University  
Indianapolis, IN 46205

Dr. Frank J. Landy  
The Pennsylvania State University  
Department of Psychology  
417 Bruce V. Moore Building  
University Park, PA 16802

Dr. Bibb Latane  
The University of North Carolina  
at Chapel Hill  
Manning Hall 026A  
Chapel Hill, NC 27514

Dr. Edward E. Lawler  
University of Southern California  
Graduate School of Business Administration  
Los Angeles, CA 90007

Dr. Edwin A. Locke  
College of Business and Management  
University of Maryland  
College Park, MD 20742

Dr. Fred Luthans  
Regents Professor of Management  
University of Nebraska-Lincoln  
Lincoln, NE 68588

LIST 15 (CONTINUED)

Dr. R. R. Mackie  
Human Factors Groups  
5775 Dawson Street  
Goleta, CA 93117

Dr. William H. Mobley  
College of Business Administration  
Texas A&M University  
College Station, TX 77843

Dr. Lynn Oppenheim  
Wharton Applied Research Center  
University of Pennsylvania  
Philadelphia, PA 19104

Dr. Thomas M. Ostrom  
The Ohio State University  
Department of Psychology  
116E Stadium  
404C West 17th Avenue  
Columbus, OH 43210

Dr. William G. Ouchi  
University of California,  
Los Angeles  
Graduate School of Management  
Los Angeles, CA 90024

Dr. Charles Perrow  
Yale University  
I. S. P. S.  
111 Prospect Avenue  
New Haven, Connecticut 06520

Dr. Irwin G. Sarason  
University of Washington  
Department of Psychology, NI-25  
Seattle, WA 98195

Dr. Benjamin Schneider  
Department of Psychology  
University of Maryland  
College Park, MD 20742

Dr. Edgar H. Schein  
Massachusetts Institute of Technology  
Sloan School of Management  
Cambridge, MA 02139

H. Ned Seelye  
International Resource Development, Inc.  
P. O. Box 721  
La Grange, IL 60525

Dr. H. Wallace Sinaiko  
Program Director, Manpower Research  
and Advisory Services  
Smithsonian Institution  
801 N. Pitt Street, Suite 120  
Alexandria, VA 22314

Dr. Richard M. Steers  
Graduate School of Management  
University of Oregon  
Eugene, OR 97403

Dr. Siegfried Streufert  
The Pennsylvania State University  
Department of Behavioral Science  
Milton S. Hershey Medical Center  
Hershey, PA 17033

Dr. James R. Terborg  
University of Oregon West Campus  
Department of Management  
Eugene, OR 97403

Dr. Harry C. Triandis  
Department of Psychology  
University of Illinois  
Champaign, IL 61820

Dr. Howard M. Weiss  
Purdue University  
Department of Psychological Sciences  
West Lafayette, IN 47907

Dr. Philip G. Zimbardo  
Stanford University  
Department of Psychology  
Stanford, CA 94305

Dr. Philip Wexler  
University of Rochester  
Graduate School of Education  
and Human Development  
Rochester, NY 14627



**END**

**FILMED**

**6-83**

**DTIC**